A Metadata Annotation Proposal (toward RT-03?)

Barbara Peskin, Elizabeth Shriberg, Jane Edwards

with contributions from:

Rowena Guevara, Steve Renals, Andreas Stolcke, Chuck Wooters

International Computer Science Institute & SRI International's STAR Lab

Objectives

Focus on “structural” information, designed for

- increased readability
- greater fluency
- improved processing by systems expecting well-formed text

Plan for near-term development, next evaluation

(was submitted in earlier form as RT-02 proposal)

What this talk is *NOT*

This is a far from comprehensive annotation scheme.

- Omits many valuable annotation types; e.g.
  - source (speaker labels; music, noise, ...)
  - “information” content (named entities, topic, ...)
- Doesn’t address *how* to mark; only *what* to mark
- Presents only first steps toward more complete framework

Where to begin?

We seek annotation types

- with good human agreement
- which provide high value for downstream processing
- common enough to be worthwhile

$\Rightarrow$ high utility, high reliability

Proposed Annotation Types

For RT-03, begin with frequent, important, (reasonably) reliable

- “utterance” units
- disruption points (later: edit intervals)
- filled pauses (later: other “fillers”)

then on to: infrequent, but helpful when occur

- quotes; parentheticals / asides

and beyond…

- prominence, back-channels, commas, ...

Utterance Units

This is the fundamental “sentence-like” unit – essential for chunking speech stream into manageable, meaningful segments

2 qualifying attributes:

- complete vs. incomplete (initial, fi nal, both)
- statement vs. question

ex: *i can’t believe he did that . what do you - ?*

ex: *right ? yeah .*

Disruption Points

Used to mark disruption of the utterance unit due to restarts, repairs, repetitions, and other disfluencies.

Represents a discontinuity for both language and prosodic models.

ex: *i'll get to it tomorr- # uh monday .*

ex: *so you really # you really believe that ?*

Next step: Edit Intervals

Used to bracket disfluent regions.
Their removal produces more “fluent” version.

Disruption points may be viewed as right-hand endpoint;
edit interval further specifies left-hand endpoint
– determine by working back from disruption point.

ex: *i’ll get to it { tomorr- # } uh monday .*
ex: *so { you really # } you really believe that ?*

Filled Pauses (later: other “fillers”)

Begin by marking standard “filled pauses”
(just “uh”, “um” in usual transcripts)

Simple token type to label (via lexical identity), hence easy step toward disfluency clean-up

Later, extend to more general “fillers”, e.g.

- “discourse markers” (*you know*, *like*, *I mean*, ...)

Other Entities for Future Work

infrequent, but reliably labelled, helpful:

- quotes

  ex: *what do you mean by “closed until further notice” ?*

- parentheticals, asides

  ex: *he responded by calling it [his words] “nuts” .*

less reliable:

- commas

  essential for disambiguating certain constructs
  (e.g. lists, certain discourse markers),
  less necessary if more structure otherwise tagged

Further Steps

Many more entities can be marked, e.g.

- prominence
- additional dialogue acts
  - back-channels, acknowledgements
  - imperatives
  - etc. etc.

All of interest, but as later stages.

How do we generate “truth”?

converting existing transcripts to tagged

- some material already marked (discussed below)
- labelling more will require effort, but...
- much can be done with simple heuristics
  ex: repetition disfluency vs. intentionally repeated word
- consistency can be increased by conventions
  ex: *yeah yeah*, *and .... and .... and ...*, *so*

Annotated Corpora

What is already labelled?

- Broadcast News – utt units (implicit in punctuation)
- Switchboard – utt units, disfluency, dialogue act
- Meetings – utt units, disfluency; dialogue acts in progress

other corpora? non-English languages?

issue: pooling sources with different annotation conventions, different information representations

Progress on Automatic Labelling

sentence (and topic) segmentation, disfluency

- BN and SWB: (Shriberg, Stolcke, Hakkani-Tur, Tur)
  June 1999 Hub 5 Workshop, Speech Communication 2000
- Meetings: (Baron, Shriberg, Stolcke)
  on-going project at ICSI

dialogue act classification

- SWB: (Jurafsky et al.)
  much work from WS97 project, but SWB not very interesting for this task
- Meetings: provide *much* richer testbed
  on-going work on statement vs. question; starting more extensive labelling

Conclusions / Proposal

- structural info markup needed for readability, downstream processing
- for RT-03, we propose starting with
  - utterance units (incl. cmplt vs. inc, '.' vs. '?')
  - disruption points (later: edit intervals)
  - basic "fillers", such as filled pauses
- high agreement, high value ...
- *We can do this!*